DOCUMENT RESUME

ED 462 445 TM 033 710

AUTHOR	Anderson, John O.; Bachor, Dan; Baer, Markus
TITLE	Using Portfolio Assessment To Study Classroom Assessment Practice.
SPONS AGENCY	Social Sciences and Humanities Research Council of Canada, Ottawa (Ontario).
PUB DATE	2001-04-00
NOTE	25p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Academic Achievement; Decision Making; Elementary Education; Elementary School Students; *Elementary School Teachers; Grades (Scholastic); *Grading; *Preservice Teachers; Student Evaluation; Student Journals; Teacher Attitudes; Teacher Education

ABSTRACT

This study focused on one task that is characteristic of teacher responsibilities and activities in the school: the evaluation of student achievement. It involved 127 preservice elementary school teachers who assessed the performance of 3 simulated students on 6 language arts tasks. Information collected included the marks assigned to students on various submitted assignments and tests and the journal entries of the student teachers. The study continues an investigation into the procedures and information bases preservice teachers use in making judgments about student achievement. The marks and grades that each student teacher generated were summarized and compared across the three simulated students to determine the extent to which the student teachers viewed their three students as distinct in their achievement in language arts. The interpretive analysis of the student teachers' journals suggests that the vast majority of these novice teachers made conservative decisions, staying close to the evidence they were given. When they had concerns, the concerns centered on their own competence or lack of background, on the appropriateness of an assignment for a particular child, and on checking to see if a student needed more help. Study findings support the view that evaluation of student achievement is not a simple process. The data show that final marks are not the same thing as final letter grades, although they are closely related. Educators have characteristic predilections to mark or grade high or low, and elements other than marks awarded to specific achievement products enter into the creation of final marks and letter grades. Results also demonstrate the potential of the portfolio approach to collecting information about the evaluation of student achievement by teachers. (Contains 11 tables and 22 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 462 445

Using Portfolio Assessment to Study Classroom Assessment Practice

John O. Anderson
Dan Bachor
Markus Baer
University of Victoria

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Anderson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM033710

BEST COPY AVAILABLE

Using Portfolio Assessment to Study Classroom Assessment Practice[1]

John Anderson, Dan Bachor & Markus Baer

University of Victoria

The evaluation of children's learning progress and achievement is a fundamental component of instruction. Over the past few years, a better understanding of how teachers conduct assessment in the classroom context has emerged (Bachor & Anderson, 1994; Broadfoot, 1992; McCallum, McAlister, Brown, & Gipps, 1992; Stiggins, Conklin, & Bridgeford, 1986). Bachor and Anderson (1994) found, for example, that teachers viewed classroom assessment as time consuming but placed a high value on 'authentic' assessment and wanted to move towards student self-assessment. Less clear, however, is how pre-service teachers develop an understanding of classroom assessment and interpret classroom assessment information. There is much work yet to be done in the area of teacher assessment practices and knowledge. The study reported here is part of a collaborative attempt to make a positive contribution to understanding teaching and learning.

The collaborative research program has been reported elsewhere (Wilson, 1999) and in several papers presented at the 2000 *Canadian Society for the Study of Education* conference (Shulha, 2000; Locke, 2000; Wilson, 2000; Petrick, 2000; Notman, 2000; Lee, 2000; Muir, 2000). This program has narrowed the focus of investigation to classroom assessment practices yet maintained a rather broad scope of investigatory approaches – including case study participant research, journal based narrative analysis, and both inferential and descriptive modelling statistical analyses – within a collaborative context (Shulha, Wilson & Anderson, 1999).

The Study

The study reported in this paper isolated its focus on one task that is characteristic of teacher responsibilities and activities in the school: the evaluation of student achievement. It involved one hundred twenty seven pre-service elementary teachers who assessed the performance of three simulated students on a number of language arts tasks. Information collected included the marks assigned to students on various submitted assignments and test, and journal entries.

The study continues an investigation into the procedures and information bases pre-service teachers use in making judgements about student achievement (Wilson & Martinussen, 1999; Shulha, 1999; Anderson, 1999). The current study utilized an evidence-based research approach to analysing the scores and grades generated by the student teachers. The portfolio structure was developed so that each portfolio contained the work of three different students on six language arts tasks, and each of the more than 100 student teachers graded the same three students. The simulated students were assumed to be in grade 5 and

[1] This study was supported by funding from the Social Sciences and Humanities Research Council Canada. The paper was presented at the annual meeting of the American Educational Rseach Association, April, 2001, Seattle, Washington.

student responses were created to reflect the work of grade 5 students. Each student teacher was required to grade each assignment as if it was requested by their sponsor teacher. Accompanying each set of student responses were instructions from the simulated sponsor teacher in regard to the grading (for example, some background to the student tasks and the total worth of each assignment). However directions in regard to how to grade the student work were designed to be rather vague and ambiguous. The student teachers were not provided with marking criteria, keys or rubrics. Student teachers were also required to maintain a journal in which they recorded the thoughts they had about the work they were doing with their portfolios. It was suggested that any comments, views, frustrations and accomplishments they encountered in marking the student work was to be noted and discussed in their journal.

The basic data layout consists of a single complex record for each participating pre-service teacher (Figure 1). Each record contains the same data elements but varied in terms of content and structure - particularly the journal entries since there was wide variation in the nature and volume of the information written by participants. The analysis of this information involved both statistical and interpretive (qualitative) approaches.

Journal Entries		**Student 1**	**marks & grades**
		Student 2	**marks & grades**
		Student 3	**marks & grades**

Figure 1: Data layout for each student teacher record

Each portfolio contained the responses of three simulated students to six language arts tasks:

1. *A Trip to the Mall* – a brief essay about going to the mall that was to be handed in as a printed word processing document. Student teachers were asked to mark this out of 12 and focus their attention on written expression rather than computer competency.
2. *Did I Order an Elephant?* – A worksheet consisting of a cloze-type reading task in which students were required to generate the 15 missing words in a reading passage.
3. *A Salmon for Simon* – A worksheet that was a modified cloze reading task in which students were required to correctly select from 5 embedded multiple-choice alternatives a phrase to complete the text, followed by 4 multiple-choice comprehension items.
4. *The New Kid on the Block* – A worksheet requiring students to read a passage and then answer 6 short answer items in which the student had to interpret a phrase from

one of the character's perspective and translate into their own words a quote from the story.

5. *The Mending Wall* – A writing task in which students had to read a 43-line poem and then write a piece describing the personal meaning they found in the poem after having developed a web outlining the main ideas in the poem and discussing this with their teacher. This was to be marked out of 25.
6. *Final Exam* – A formal written exam consisting of 20 word identification items (classify word as a noun, adjective, verb or adverb), a paragraph in which the student had to extract 5 nouns and 5 verbs, 14 editing items for commas and correct capitalization, and a reading passage followed by 8 multiple-choice comprehension items and 2 written responses (5 and 10 lines of space provided for student response).

For each of these tasks, three responses or achievement products were created for inclusion in the portfolio. One product was developed to represent low achievement, another was created to represent mid-level achievement, and one was created to represent high achievement. The development of the achievement products (simulated student responses) involved both the researchers and some elementary school students who developed and located the level of responses. The low-level products were consistently assigned to *Student A*, the mid-level products assigned to *Student C* and the high-level products assigned to *Student B*. The effect intended was that each student teacher would assemble a portfolio of achievement products for three students: a high achiever, a moderate achiever and a low achiever.

Table 1: Summary Statistics for Marks Awarded Students A, B & C.

TASK	STUDENT A (Low) Mean (SD)	B (High) Mean (SD)	C (Mid-level) Mean (SD)
Trip to Mall	8.0 (1.5)	9.5 (1.3)	10.9 (1.2)
Did I Order an Elephant?	5.9 (0.8)	7.6 (0.6)	6.6 (0.8)
Salmon for Simon	5.1 (1.0)	7.4 (0.8)	5.1 (1.0)
New Kid on the Block	13.2 (2.6)	16.3 (1.7)	14.4 (2.3)
The Mending Wall	12.6 (3.2)	23.2 (1.6)	20.0 (2.4)
Final Exam	22.6 (3.6)	46.2 (3.7)	37.6 (3.2)

The summary statistics (Table 1) and the plots (Figure 2) show that this design expectation was realized for most of the achievement products in that Student A was awarded the lowest scores, Student C mid-range scores and Student A the highest scores. There were two exceptions to this pattern. One exception was *A Trip to the Mall*, the first product given to the student teachers for inclusion in the portfolio where Student C (the moderate achiever) was generally awarded higher marks than our high achiever (Student B). The other exception was the similar results for Student A (low) and Student C (mid) on their results for *A Salmon for Simon* worksheet. There were no tasks on which the average scores of the low achiever (A) were greater than those of the high achiever (B). On the basis of these descriptive findings it was concluded that the achievement products included in the assembly of the portfolio were representative of low, moderate and high achieving students.

The Statistical Analysis

The marks and grades that each student teacher generated were summarized and compared across the three simulated students. The goal was to investigate the extent to which the student teachers viewed their three students as distinct in terms of their achievement in language arts. The design of the portfolios was intended to create a low achieving student, one that was high and another who was a mid-range achiever through the development and inclusion of student work that consistently represented what was viewed as low, mid and high ranges of achievement. The extent to which these results are reflected in the grades and marks assigned by the student teachers could be considered an index of the design representativeness of the portfolios.

The intercorrelations of marks and grades were calculated to investigate the extent to which each assignment and test yields the same kind of information about the simulated student. Since the underlying factor in the student work is language arts achievement, it was anticipated that strong, positive correlations within each student's set of marks would emerge.

Student teachers were requested to mark each of the six achievement products of each of the three students and then submit a final mark and lettergrade for each student. Eighty-two student teachers submitted both a *Final Mark* (a numerical score) and a *Lettergrade* (104 submitted a *Final Mark* only). For analysis purposes the *Lettergrades* were transformed into numbers with A being given a value of 7, an F (fail) being assigned a 1 and the rest of the grades ranged accordingly in between.

In the marking of two of the six assignments, student teachers varied a bit in the maximum scores allowed by the sponsor teacher. For the task *Did I Order an Elephant?*, maximum scores ranged from 5 to 45 with most student teachers using a maximum of either 8 or 15. For the *Final Exam* maximum scores ranged from 41 to 82 with most student teachers using a maximum of 50. For analysis purposes all scores were transformed to a common scale: *Did I Order an Elephant?* was scaled to a maximum of 8, and the *Final Exam* was scaled to a maximum of 50.

Figure 2: The Results for Students A, B & C on the Tasks

Figure 2a Results Trip to the Mall Task

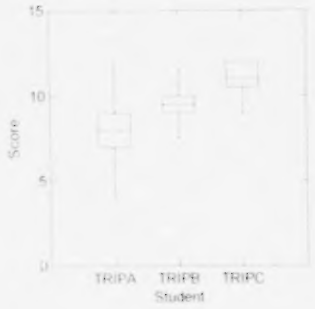


Figure 2b Results Did I Order an Elephant?Task

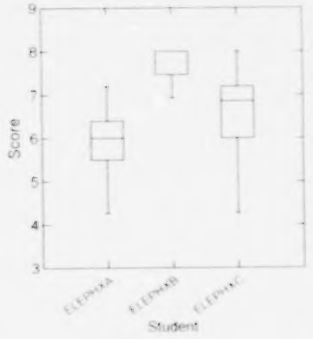


Figure 2c Results for Salmon for SimonTask

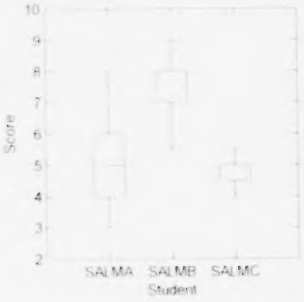


Figure 2d Results for New Kid on BlockTask

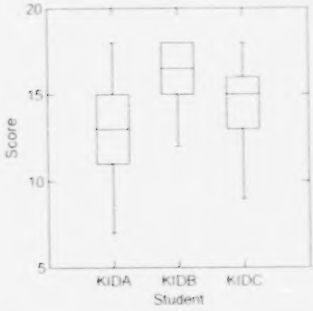


Figure 2e Results for The Mending Wall Task

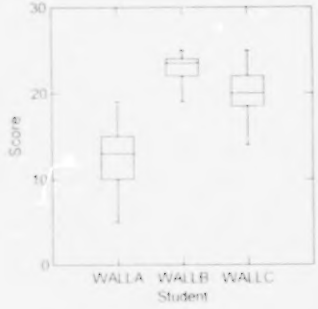
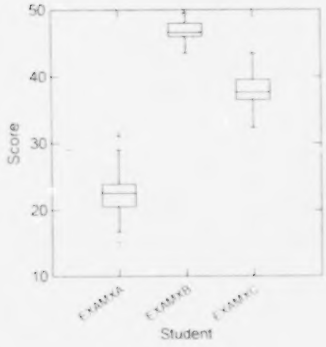


Figure 2f Results for Final Exam



In considering the summary statistics it is apparent that student teachers assigned a wide range of scores to the three students on the various tasks included in the portfolio. For any particular task there is overlap in the scores assigned students A, B and C. However the ranking of the three students who were simulated in the portfolio was consistent: crosstabultation of the final lettergrades indicated that Student B was always rated first in achievement, Student C second and Student A was consistently ranked last. A distribution of lettergrades (Table 2) shows considerable consistency across student teacher evaluations with Student B having a modal grade of A, Student C having a mode of B and Student A a modal grade of C⁻ - results which are consistent with the design of the portfolio contents. However there is substantial variation in the grades awarded by different student teachers – for example, five student teachers awarded the highest achieving student a grade of B, the grade awarded to the lowest achieving student by three other student teachers.

Table 2: Lettergrade Distributions

Student	Grade						
	F	C⁻	C	C⁺	B	A	A⁺
High (Student B)	-	-	-	-	5	73	4
Moderate (Student C)	-	1	-	8	66	6	-
Low (Student A)	10	41	23	4	3	-	-

Correlations were calculated for all marks (Table 3) to investigate relationships between achievement tasks and among students. Preliminary perusal of the correlations suggests that there are higher correlation between instruments than between students. For example, the correlations between *Trip to the Mall* for students A, B and C are 0.56, 0.27 and 0.47 whereas the correlation between marks awarded to a given student on this assignment are all lower than 0.27. This pattern is typical of the correlation structure for these results and indicates that a student teacher who awards a high student score on one assignment will tend to award a higher score for students on the assignment. In other words, a high score awarded to Student A on assignment one would be related to a high score awarded by that student teacher to Student B on that same assignment. This *marker tendency* was further revealed by the correlations of final marks for each of the three students (Table 4). The correlations are all positive ranging form 0.21 to 0.47 suggesting that student teachers who award a high final mark to Student A will also tend to award a high final mark to Students B and C, whereas a student teacher awarding a low final mark to Student A will tend to award a low final marks to Students B and C.

Table 3: Correlations between achievement products for the three students

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1 TRIPA	1.00																
2 TRIPB	0.56	1.00															
3 TRIPC	0.27	0.47	1.00														
4 SALMA	0.05	0.02 -	0.11	1.00													
5 SALMB	-0.03	0.05	-0.10	0.20	1.00												
6 SALMC	0.02	-0.09	-0.13	0.56	0.08	1.00											
7 KIDA	-0.05	0.04	-0.21	0.22	0.15	0.07	1.00										
8 KIDB	0.02	0.15	-0.06	0.21	0.12	-0.05	0.30	1.00									
9 KIDC	0.16	0.06	-0.09	0.16	0.01	0.03	0.35	0.35	1.00								
10 WALLA	0.12	0.03	0.00	-0.05	0.02	-0.09	-0.02	-0.00	0.12	1.00							
11 WALLB	0.10	0.04	-0.07	0.12	0.16	0.04	0.10	0.10	0.11	-0.02	1.00						
12 WALLC	0.19	0.09	0.05	-0.05	-0.02	-0.03	-0.14	0.04	0.21	0.29	0.39	1.00					
13 LEPHXA	0.09	0.22	-0.01	0.18	0.05	0.04	0.02	0.06	0.05	0.05	0.13	0.06	1.00				
14 LEPHXB	0.24	0.22	0.09	0.07	0.14	0.06	0.20	0.15	0.13	-0.00	0.29	0.21	0.28	1.00			
15 LEPHXC	0.22	0.17	-0.00	0.18	0.09	0.09	0.16	0.11	0.19	-0.11	0.34	0.09	0.36	0.65	1.00		
16 LXAMXA	0.06	0.08	0.15	0.05	0.26	-0.04	0.13	0.15	0.19	0.18	0.01	0.21	0.03	0.25	0.04	1.00	
17 LXAMXB	0.13	0.13	0.01	0.15	-0.01	0.08	-0.01	0.00	0.01	0.00	0.15	-0.00	0.12	0.21	0.29	-0.22	1.00
18 LXAMXC	0.07	0.06	0.10	0.03	-0.02	0.11	0.10	0.04	0.04	0.15	-0.15	-0.00	0.12	0.12	0.07	0.30	0.25

Table 4: Correlations of Final Marks

Student	A (Low)	B (High)	C (Moderate)
A			
B	0.21		
C	0.47	0.37	

As it turned out, although there was a strong positive relationship between the final mark and the lettergrade awarded it was nowhere near perfect (Table 5). The final mark accounted for 38 to 76% of the variance in the lettergrade awarded, depending on Student. Since it was of interest to explore the nature of

Table 5: Correlation between *Final Marks* and *Lettergrades*

	Correlation	r^2
Student A	0.87	.76
Student B	0.62	.38
Student C	0.79	.62

the elements contributing to the final marks and lettergrades awarded to the three students by the student teachers regression analyses were conducted using the six achievement products to predict the final mark and the lettergrade for each of the three students in the portfolio. The results show that the *Final Marks* (Table 6) are better accounted for (R^2's range from 0.80 to 0.87) than *Lettergrades* (Table 7 where R^2's range from 0.31 to 0.66). The final results, particularly for lettergrades, are in some way constructed differently for each of the three students. For example, the 66% of the variance of *Lettergrades* awarded Student A (the low achieving student) are accounted for by the marks awarded the six achievement products whereas for Student B (the high achiever) only 31% of *Lettergrade* variance is accounted for by the marks and almost 70% is from other sources.

Table 6: Regressions of Marks on Achievement Products to *Final Mark* for Students A, B and C

Achievement Product	Coefficient (β)			p			R^2		
	A	B	C	A	B	C	A	B	C
Trip to Mall	0.20	0.17	0.15	.00	.00	.01			
Salmon for Simon	0.22	0.12	0.14	.00	.03	.01			
New Kid	0.35	0.28	0.36	.00	.00	.00			
Mending Wall	0.44	0.30	0.38	.00	.00	.00			
Elephant	0.17	0.10	0.17	.00	.07	.00			
Final Exam	0.50	0.63	0.55	.00	.00	.00			
							0.89	0.80	0.80

Table 7: Regressions of Marks on Achievement Products to *Lettergrade* for Students A, B and C.

Achievement Product	Coefficient (β)			p			R^2		
	A	B	C	A	B	C	A	B	C
Trip to Mall	0.22	0.02	0.04	.01	.87	.68			
Salmon for Simon	0.15	0.05	0.09	.06	.65	.35			
New Kid	0.19	-0.04	0.17	.02	.73	.09			
Mending Wall	0.41	0.34	0.34	.00	.00	.00			
Elephant	0.18	-0.03	0.18	.03	.81	.06			
Final Exam	0.33	0.34	0.40	.00	.00	.00			
							0.66	0.31	0.51

In summary it can be said that the *Final Mark* (the numerical final result) is well accounted for by the marks awarded on the six achievement products, and that each of the six achievement products contributed significantly to the *Final Mark*. Further, the relationships of achievement products to *Final Mark* are consistent across three students. However for the *Lettergrade*, the final results are not as well

accounted for by the six achievement products. Further the extent to which achievement products contribute to the final result varies one student to another and only two of the six achievement products (*Mending Wall* and the *Final Exam*) significantly contribute to the final lettergrade. This suggests that the numerical version of the final result is not quite the same thing as the *Lettergrade* although conceptually they should convey the same information.

The Interpretive Analysis

The pre-service teachers were asked to keep a journal in which they were to record their comments about their assessment process. The journal entries were explored and analyzed to reveal elements and patterns in the thoughts, concerns and issues that student teachers expressed as they were attempting to complete their task of grading their three students. The student teachers were all given the same materials on the students, the cooperating teacher and the school. Since this information was rather sparse, there were likely to be variant interpretations of the task and situation. Ambiguities of expectations and task definitions were issues that were expected to be expressed in the journals. As well, the rather limited information provided on each of the three students in the portfolio has created a more decontextualized evaluation situation than what is likely to occur in most classrooms. It is expected that the extent to which this is noted as an issue in the journals may be a major element of the journal data. However, the analysis of the journal entries provided a rich source of information about the concerns and thoughts related to the evaluation of student achievement.

These journals served as the data for the interpretive analysis. The journal entries themselves varied in length from several lines to numerous pages. The original journals were transcribed, translated into a 'text file', and then stored as a single 'primary document' as an *Atlas/ti* (Muhr, 1997) file. The journal data was then analyzed for patterns.

Preliminary Coding. As a starting point, preliminary codes were developed from informed practice and the assessment literature (eg., Bachor & Anderson, 1994). After the establishment of these initial categories, following Glaser and Straus's 'constant comparison' method (Tesch, 1990), data from the first three participants was repeatedly coded with the goal of refining and reestablishing codes. Following the establishment of these preliminary codes, data from the first three cases were coded several times to 1) verify that the codes could be consistently applied across cases by both authors, 2) ensure that codes were comprehensive enough to allow the evidence to be classified comprehensively, and 3) ascertain that the codes did not contain redundancies.

Although the codes have their origins in existing theory and practice, they are grounded in the data to accurately and comprehensively represent the journal entries. A secondary purpose of repeated coding and comparison was to train for consistency. The relatively 'open ended' nature of the diary task resulted in responses that were at times vague or ambiguous. Thus, code category boundaries required revision and refinement in order to deal with textual uncertainties. In turn, redundancy and overlap between categories was reduced.

Code development. Based on the literature and an initial examination of the *Atlas ti* data, three superordinate categories were identified (Table 8). Initially, comments were divided into those that were primarily "assignment based" (dealing with the context, of the work completed, responding to the assignment criteria, or reacting emotionally to the assignment itself) and those labeled "person based" (describing the competency, quality of life or other comments directed specifically at the theoretical student as a person). Subsequently, a third category termed "intervention" was added to parse out intervention suggestions, taking the form of either comments or directives aimed at specific students. The three core codes of "assignment based", "person based" and "intervention" proved to fit the data upon subsequent re-workings of subordinate categories. Eliminating redundancy, overlap and ambiguity in lower order code categories required several further revisions before fourteen final codes were established. The final fourteen codes classified into three superordinate categories are given in Table 8.

Journal entry ambiguity. Despite reworking the codes to reflect and adequately represent the complexity of the journal data, ambiguity and vagueness in the language of some participant journal entries remained. For example, regarding one student's assignment, a participant wrote, "Watch for comprehension in other areas". It is unclear whether the comment is a reminder to the teacher/participant, or a word of advice - suggesting an intervention - to the student. In another example, a participant wrote, "Student needs to work on context of her statements". Again, it is uncertain whether this suggests an intervention, merely advises the student where they erred, or is simply an effort to justify the grade assigned for the task. In such cases, face validity of the text was assumed and comments were taken at the textual level. The large data set rendered verification of codes with participants impractical, and thus, textual inferences were kept to a minimum. Lower order or broader code categories were applied when there was uncertainty. In the both of the above cases, for example, the comments were coded with the larger category of "person based competency-performance on task".

Inter-judge agreement. Reliability checks for the code categories were conducted. Two of the researchers independently coded three randomly selected sections of text consisting of between 100 and 150 lines per section on two separate occasions. A random number table was used to select the text segments. The independently coded sections were compared for consistency of code application using point by point agreement ratios (Kazdin, 1982). Reliability rates were checked twice: for the first check was 72%, and for the second one was 96%. The average reliability rating was 78% agreement.

Using the categories given in Table 8, codes were applied to the collected text of all 127 participant diaries. The amalgamated data was treated as one primary document and coded in its entirety prior to any analysis. Upon completion, participants were each given their own code in order to examine differences both across and between this group of pre-service teachers. Throughout the data entry, we met to check for coding agreement and to ensure consistency of coding.

Table 8
Codes Assigned to Participants' Diary Data

Code	Definition
	Superordinate Category
	Assignment-Based
Context-	
Classroom	**Points raised about the task, teacher, classroom, et cetera**
Subject's Background	**Comments made about the pre-service teacher's own background**
Criteria-	
Establishing	**Process of establishing assessment criteria**
Reviewing/ Refining	**Subsequent reviewing and refining of initial criteria**
Questions Comments-	
Concerns	**Queries raised about the assignment/task**
Positives	**Comments made about the assignment/task**
	Intervention
Comments	Hints of an intervention, such as suggestions directed at task, class, teacher, et cetera
Student	Specific suggestions for an intervention, directed at either student A, B, or C
	Person Based
Competency-	
Performance on Task	**Statements about performance on task, directed to Student A, B, or C indicating how well he/she did on an assignment**
Student	**Statements directed at the student going beyond task comments, designating the student, eg. Student A is poor speller**
Classification	**Statements directed at the student going beyond assignment comments. Designating one of the students as having a special educational need, eg. Learning Disabled, gifted, et cetera**
Quality of Life	**Statements directed at the student's family, such as commenting about their social economic status**
Comments-	
Knowledge of	**Comments indicating that knowing the student was important to participant's understanding of his/her progress as a learner**
Affective State	**Statements made about the emotional state of either Student A, B, or C**

Results - Interpretive Analysis

In presenting these results, we began with the evidence collected from all 127 participants and then we parsed the data into a number of different groupings based on the conclusions that we deduced the participants made. Two main distinctions were drawn. First, we isolated those individuals who we called *Task Restricted Participants* (TRP). Second, we pinpointed a second small cluster of participants, whom we named *Student Elaboration Participants* (SEP). Based on the comments that they made in their journals, novice teachers tended to follow one of two main decision paths in interpreting children's assignments. The majority of PTs, the Task Restricted Participants, seemed to be quite conservative in the decision path they appeared to follow (Table 9). However, a minority of individuals, the Student Elaboration Participants (SEP) - appeared to make extreme decisions regarding the hypothetical students they assessed (Table 10). In reading these tables, note that we have progressively eliminated an increasing number of participants as we describe the factors that individuals seemed to consider when making decisions. For example, in Table 10, we begin by presenting the decision-path of all 16 SEP, thus the reduction in number of participants noted above. As you read down the table, progressively more assessment-comments — 'quality of life' and 'affective state' in the first instance — are added to note the decreasing number of SEP, who included other factors in their decision-making about the three hypothetical children.

The vast majority of participants (Table 11) established some criteria to judge the assignments they received (124 out of 127 participants). To illustrate, typical comments by PTs are the following two, where the focus is establishing guidelines for marking:

> "Each response is out of 3. There are 6 questions so task is out of 18 marks. 1 mark is given for each criteria (sic): - is idea relevant to story & character 1- express ideas as Jimmy (I or me) 1 - sentence thoughtful & clearly expressed 1".
>
> "Basically, I marked the answer correct if it seemed to reasonably fit into the context of the sentence. Although there were several instances where one student gave a much more appropriate response than another, I marked both of them right because they both were reasonable answers."

Some individuals elaborated the criteria they proposed, commenting extensively about the assignment they were assessing. For example, one person noted,

> "As I marked this assignment, I specifically looked for reading and writing comprehension. I read each student's answer in context with the sentence and the story. In Part 2 I had trouble deciding what was the right answer for #3. I kept marking it wrong then right, so I decided to give everyone a mark for their answers. I do believe that Student B's answer was the most thought out and appropriate, but I also saw how Student A and C might have interpreted the questions and answered accordingly. Each answer was marked out of 1 mark for a total of 9 marks."

Table 9
Task Restricted Participants' Decisions

Decision Path	Number of Participants
All Participants	127
Participants developing criteria ('Assignment Based Criteria-establishing')	124
Participants making 'Person Based Competency-performance on task' statements	105
Excluded Participants	
Participants who made 'classification' comments	16
Participants who made 'quality of life' comments	17
Task Restricted Participants (TRP)	
Participants who made no 'classification' or 'quality of life' comments	100
TRP who made no 'affective state' comments and no 'Intervention' statements	
TRP who made no 'affective state' comments, and no 'Intervention-comments'	52
TRP who made no 'affective state' comments, and no 'Intervention-student' comments	56
TRP who made no 'affective state' comments, and no 'Intervention-comments' or 'Intervention-student' comments	33
TRP who made no 'affective state' comments and no 'Intervention' statements or 'Person Based Competency-student' comments	
TRP who made no 'affective state' comments, and no 'Person Based Competency-student' comments	46
TRP who made no 'affective state' or 'Intervention-comments' and no 'Person Based Competency-student' comments	27
TRP who made no 'affective state' or 'Intervention-student' comments and no 'Person Based Competency-student' comments	36
TRP who made no 'affective state', 'Intervention-comments' or 'Intervention-student' comments and no 'Person Based Competency-student' comments	22

Note. All numbers refer to participants who coded at least once or more with the specified categories.

A small number of novice teachers (13/127 participants, Table 11) were not satisfied with the initial criteria they established. They revisited the criteria they established, either prior to or during the process of assessing assignments. For example, one person noted

> "This is a rather difficult assignment. I wasn't even sure of some answers. As such, I modified my original marking scheme. I started out thinking that it would be smart to mark the first 5 either right or wrong, but I ended up giving 1/2 marks if it was semi-relevant, 0 if not consistent with the story, and 1 for the best choice. That way, the marks weren't so low."

Table 10
Student Elaboration Participants

Decision Path	Number of Participants
All Participants	127
Student Elaboration Participants (SEP) ['Person Based Competency-classification']	16
SEP who made 'quality of life' comments	6
SEP who made 'affective state' comments	8
SEP who made 'quality of life' and 'affective state' comments	5
SEP who made 'Intervention' statements	
SEP who made 'Intervention-comments'	11
SEP who made 'Intervention-student' comments	10
SEP who made 'Intervention-comments' and 'Intervention-student' comments	7
SEP who made 'Intervention-comments' OR 'Intervention-student' comments	14
SEP who made 'quality of life', 'affective state' and 'Intervention-comments'	4
SEP who made 'quality of life', 'affective state' and 'Intervention-student' comments	4
SEP who made 'quality of life', 'affective state', 'Intervention-student' and 'Intervention-comments'	3
SEP who made 'quality of life', 'affective state', 'Intervention-student' OR 'Intervention comments'	5

Note. All numbers refer to participants who coded at least once or more with the specified categories.

In addition, 50 PTs (Table 11) made comments about the context of the assignments they were asked to assess. These remarks centered on the artificial nature of the assessment, as the PTs were not setting the assignments but were judging work given by a hypothetical grade 5 teacher, who is not well described in the context of the study since the focus is on the three hypothetical students. For example, one novice teacher commented "Because I do not know exactly what the teacher has discussed with the students before doing the assignment it is more difficult to mark on what they actually wrote about (content)", while another was concerned about previous student learning, writing "I wonder if students have worked with poetry before. I hope so cause this is a heady poem to interpret". Further, 21 individuals expressed discomfort in assessing some components of the assignments given due to weaknesses in their own background. For example, one person noted "Because I do not have much experience with marking I tend to question what I am doing".

Table 11
Participant Count by Code Categories

	Participant Count
Assignment Based Context-classroom	50
Assignment Based Context-subject's background	21
Assignment Based Criteria-establishing	124
Assignment Based Criteria-reviewing/refining	13
Assignment Based Questions/Comments-concerns	112
Assignment Based Questions/Comments-positives	71
Intervention-comments	56
Intervention-student	59
Person Based Competency-classification	16
Person Based Competency-student	74
Person Based Competency-performance on task	105
Person Based Quality of Life	17
Person Based Student Comments-affective state	28
Person Based Student Comments-knowledge of	21

Note. Participant count includes all participants who contained one or more instance of the specified category.

Task Restricted Participants.

Characteristics. As depicted in Table 9, the vast majority of individuals (100/127), whom we term task restricted participants (TRP), did not make any comments beyond judging the hypothetical children's work. That is, they tended to confine their comments those related to the assignments, such as establishing criteria, without making any classification or quality of life comments regarding the student personally.

Excluded individuals. Twenty-seven participants (Table 9) were excluded from further analysis in this category because they did not meet the criteria for task-restricted. Of this total, 17 participants made quality of life statements and 16 participants concluded that some children had special educational needs. There was an overlap between these two sets of comments, as 6 participants made both types of statements. Some of these individuals will be examined latter under the category of student-elaboration.

TRP Patterns. Examining Table 9 reveals that some of the task-restricted participants were very cautious as to the statements they made. Some individuals (52/127) did not make any comments regarding the children's affective state (affective state comments ranged from neutral statements about not wanting to hurt a child's feelings to ones indicating that a child was unhappy at school), nor did they make any general intervention suggestions (such as an assignment may need to be rethought). Fifty-six TRP did not make

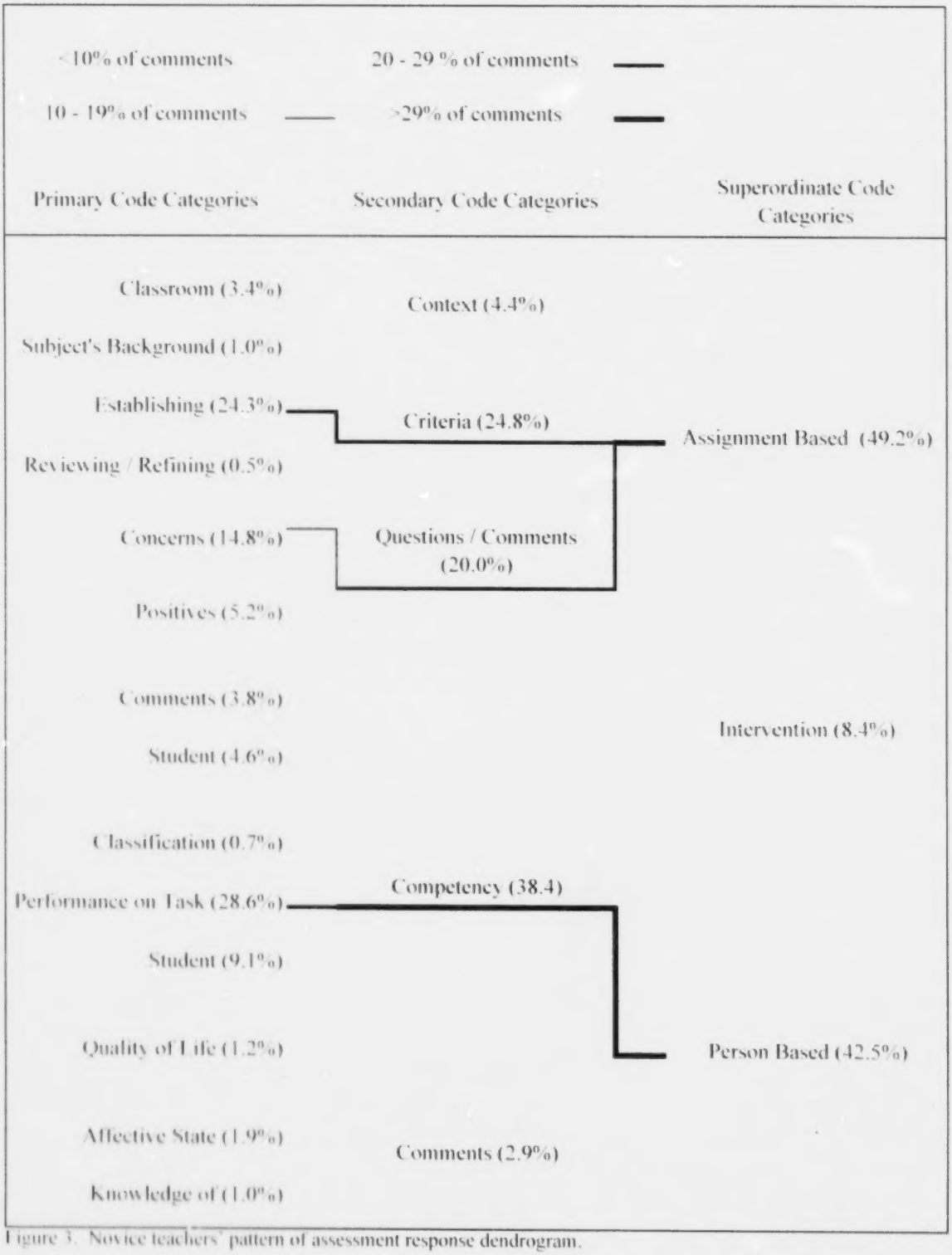


Figure 3. Novice teachers' pattern of assessment response dendrogram.

affective state comments or make student-specific intervention comments, such as Student A needs help in spelling. A smaller sub-set of participants (33) made neither type of intervention statement, nor did they offer affective state comments.

The most conservative group (22/127), in addition to following the above pattern, further restricted their comments. They did not make any judgements about individuals general abilities (Person-based Competency student), such as a student can not spell. As depicted in the last part of Table 9, there were other variations on this pattern of the type of comments made.

Student Elaboration Participants

A small number of individuals, referred to as *Student Elaboration Participants* (SEP), however, appeared to be willing to make judgements that exceeded the evidence provided. Seventeen novice teachers (Table 11) made quality of life comments; that is, they commented about the quality of the family home or student's social life and how it was thought to have influenced the hypothetical child's school performance. For example, one SEP stated "Student A seems to have a poor family life and it's reflected in his her work." Another small group made comments about the affective state of some of the hypothetical students they were assessing. An example of this kind of comment is as follows: "Hard worker and likes to do many things at once – I'm hoping this won't be a detriment (pressure >stress)."

Additional substantial judgements were made by 16 SEP who were willing to designate one of the children as having special educational needs based on very limited evidence. They made comments like "I'm wondering if they are ESL or some type of learning disability – why didn't the teacher offer extra assistance at some point?". Six of the 127 SEP also made quality of life comments, while a further subset of 5 of the 6 SEP made designations not only of special educational needs and quality of life concerns, but also went on to offer an intervention directed at the student. For instance, one SEP commented

> "Student A needs a great deal of work with grammer (sic), spelling and sentence structure. I am wondering if this student has a learning disability or not one of the greatest home lives... This student needs a great deal of encouragement and assisstance (sic). I hope that s/he gets it."

Interventions

Limitation. While we were able to isolate comments made by SEP, we were not able to completely differentiate between the types of intervention statements made across participants. Thus, there might have been some overlap in interventions comments offered by the various PTs.

Intervention-comments. General comments about assignments (Table 11) were made by 56 of the 127 participants. Prototypical examples include the following three, illustrate the range of comments made.

> One individual suggested, "As follow up, I would ask students to re-read their work for structure problems and make a lesson out of it".
>
> Focusing on the educator's role in the assignment, a PT stated, "The teacher should go over components/characteristics of an essay – paragraph breaks indentations etc.".

Another PT commented on what they themselves might do, relating that "I would spend much time reviewing this sheet because the students obviously did not understand this concept. Also a follow up lesson was needed to ensure it was learned as was done.".

Intervention-students. In addition, 59 PTs made intervention comments specific to one of the three students. Examples of this latter type include the following three.

- One PT noted some additional work might be required in rethinking an assignment. "I would perhaps return student A's paper and let him/her redo the assignment".
- Another person suggested that one of the hypothetical students might need some assistance in writing "She need to work on her run-on sentences; look out for these in the future".
- Finally, another PT offered suggestions to improve spelling. "I would encourage the student to use the dictionary and read over and proofread work for errors. Student may also have a peer read or assist with spelling. I would also encourage the student to slow down when he writes & try to write on the lines. I may have the student complete grammar exercises".

Reframing the Evidence: A Dendrogram

Examining the evidence from another perspective, the PTs comments about the assessments were divided into two main categories. Looking at the dendrogram given in Figure 3, approximately one-half (49.2%) of the diary entries focused on the assignments the PTs addressed. These were divided into two main sub-groups: setting or reviewing criteria (24.8%) and asking questions or commenting on the assignments (20%). The second common cluster of comments was centered on the three hypothetical students (42.5%). As can be seen, the bulk of these comments (38.4%) focused on the hypothetical students competency. The large majority (28.6%), however, were restricted to addressing specific aspects of the children's performance on the language arts assignments. A minority of comments (eg. quality of life, 1.2%, or classification statements, 0.7%), however, were not supported by the evidence provided in the portfolios.

Discussion

A key limitation in analyzing the diaries of the novice teachers who took part in this study is that we were not able to verify that the comments made actually reflect the decisions that these PTs would make in the classroom. Each person was asked to comment about the process they were following as they assessed the three hypothetical students and we took these comments at face value. In addition, it is important to interpret our findings with caution since even those participants who made seemingly extreme comments often added contextual qualifications to their remarks. Thus, we can not ensure that the decision-paths that we traced were the specific ones taken by the various participants.

The vast majority of the novice teachers in this study appeared to make conservative decisions, staying close to the evidence they were given over the course of an academic term. When they had concerns, they centered on their own competence or lack of background, on the appropriateness of an assignment for a particular child, or on checking to see if a particular student needed some additional help

in mastering some aspect of language arts. These individuals are consistent with the assessment patterns demonstrated many other teachers (see Shuhla, 1999, for example). Following the scheme suggested McCallum, McAlister, Brown, and Gipps (1992), these novice teachers seem to be becoming Systematic Planners. That is, the majority of PTs appear to be developing into teachers who systematically incorporate assessment evidence into their teaching practice.

A minority of individuals, however, presumably made assessment decisions that far exceeded the evidence provided. They seemed prepared to base their assessment decisions on some undefined assumptions. They appeared to have an intuitive basis for the judgements they made and speculated willingly about the three hypothetical learners and their families. Others (eg., Bachor & Anderson, 1994; Broadfoot, Abbott, Osborn, Pollard, & Croll, 1993; Stiggins, 1999) have also noted the idiosyncratic nature of assessment. They have urged teachers to be prudent and systematic when conducting classroom assessment, as the cost of teachers using unsound assessment practices is too high.

Teacher educators can take some comfort in knowing that novice teachers, for the most part, have the skills to make fair assessment decisions and appear to be making reasonable decisions. One unanswered question, however, is whether these competencies will be utilized in the classroom context where teachers have different levels of commitment to the students that they are interacting with on a daily basis. In the present case, their presumed impartiality of the majority of participants may be a reflection of judging hypothetical students or other unidentified considerations.

For a small number of novice teachers, teacher educators must be very vigilant in addressing the assumptions that seem to be held by any individuals who are prepared to make judgements based on sparse evidence. This concern is particularly justified when we consider the larger context of teachers' classroom assessment decision-making. Previously concern has been expressed over the basis that some teachers use to make decisions (eg. McCallum, McAlister, Brown, & Gipps, 1992). Specifically, some teachers make decisions about children based on their intuitive sense of a child, on the family and school history, or on very limited encounters with an individual. These decisions tend to become rigid and are subsequently not readily amended. Whether teacher educators can influence such individuals to shift their assessment practices is unknown; however, every effort must be made to redress unsound assessment practices.

In Closing

The study supports the view that the evaluation of student achievement is not a simple process. The numerical data shows clearly that final marks are not the same thing as final lettergrades although they are closely related. Educators have characteristic predilections to mark or grade high or low — *marker tendency* — which we believe corresponds to many students' recollections of grades past. Further, elements other than the marks awarded to specific achievement products (worksheets, assignments and tests) enter into the creation of the final marks and lettergrades teachers assign to students, and more additional information is added into the creation of lettergrades than into numerical final marks. And finally, that the information used in the composition of lettergrades varies one student to another.

The results indicate the potential for the portfolio approach to collecting information about the evaluation of student achievement by teachers. The achievement products created for this portfolio appear to have functioned in the manner intended in that the low achieving student was perceived to be low, as did the high achieving student and the moderately achieving student.

The next steps in this research will focus on these as yet unknown information elements teachers used to develop their grades for students. To do this the information written by the student teachers in their journals should provide insightful. The patterns revealed in the interpretive analysis of the journals will be used to investigate the structures underlying the lettergrades assigned to our three students of varying achievement. Since the journal entries are linked to the marks and grades the patterns in the journal data can inform the further analysis of the marks and grades. Categorizing meaningful patterns found within the journal data will allow for the use of both teacher perspectives from the journal entries and assigned marks in statistically modeling the evaluation of student achievement which will constitute the next stage of the investigation. This will prove to be a complex task. The use of journal entries for the development of categorical information will be fed into a structural equation model, this should allow for the development of a model that is based upon structures suggested by the thinking of the individuals generating the achievement data. The previous studies in this research embedded information about the simulated student into the portfolio materials. The model developed from these data (Anderson, 1999) was meaningful but accounted for a relatively small proportion of variance in the assessment data. It is anticipated that the results of these future analyses based on the data reported in this paper should provide a basis to the development of a model that will facilitate the study of the structures underlying the evaluation of student achievement.

References

Anderson, J.O. (1999). Modeling the development of student achievement. *Alberta Journal of Educational Research*, 45(3), pp. 278-287.

Bachor, D.G. & Anderson, J.O. (1994). Perspectives on Assessment Practices in the Elementary Classroom in British Columbia, Canada. *Assessment in Education: Principles Policy and Practices*, 1, 65-95.

Broadfoot, P. (1992). Assessment developments in French education. *Educational Review*, *44*, 309-326.

Broadfoot, P., Abbott, D., Osborn, M., Pollard, A. and Croll, P. (1993, September). SATS and Teacher Assessment: The Changing Reality of National Assessment in Primary Schools. Paper presented at the British Educational Research Association Conference, Liverpool.

Kazdin, A.E. (1982). *Single Case Research Design: Methods for Clinical and Applied Settings*. New York: Oxford University Press.

Lee, M. (2000). *The emergence and anatomy of collaboration in a teacher-researcher assessment project: A cognitive and cultural view*. A paper presented at the annual meeting of the Canadian Society for the Study of Education, University of Alberta, May, 2000.

Locke, C. (2000). *Student control and responsibility: The influence of two teachers' classroom assessment*. Paper presented at the annual meeting of the Canadian Society for the Study of Education, University of Alberta, May, 2000.

McCallum, B., McAlister, S., Brown, M., & Gipps, C. (1992, August). Teacher assessment at key stage one. Paper presented at the British Educational Research Association Conference, Stirling.

Muir, W. (2000). *Collaborative research and the internet: A case study*. Paper presented at the annual meeting of the Canadian Society for the Study of Education, University of Alberta, May, 2000.

Muhr, T. (1997). *Atlas.ti: The Knowledge Workbench*. Version 4.1, Berlin Scientific Software Development.

Notman, D. (2000). *The effects of portfolio assessment and student-led conferences on ownership and control*. Paper presented at the annual meeting of the Canadian Society for the Study of Education, University of Alberta, May, 2000.

Petrick, B. (2000). *The relationship between learning and assessment in the elementary classroom*. Paper presented at the annual meeting of the Canadian Society for the Study of Education, University of Alberta, May, 2000.

Shulha, L. M. (1999). Understanding novice teachers' thinking about student assessment. *Alberta Journal of Educational Research*, 45(3), 288-301.

Shulha, L. M. (2000). Implementing newer principles of assessment into planning and delivery of instruction. Paper presented at the annual meeting of the Canadian Society for the Study of Education, University of Alberta, May, 2000.

Shulha, L. M., Wilson, R.J. & Anderson, J.O. (1999). Investigating teachers' assessment practices: Exploratory, non-foundationalist, mixed method research. *Alberta Journal of Educational Research*, 45(3), 304-314.

Shulha, L. M., Anderson, J.O., Muir, W., Wilson, R.J., & Bachor, D. G. (1997, June). Collaboration in Research and Development. Paper presented at Canadian Society for Studies in Education Conference, St. Johns, Newfoundland.

Stiggins, R.J. (1999). Evaluating Classroom Assessment Training in Teacher Education Programs. *Educational Measurement: Issues and Practice*, 18, 23-27.

Stiggins, R. J., Conklin, N.F. & Bridgeford, N. J. (1986). Classroom assessment: A key to effective instruction. *Educational Measurement*, 5, 5-17.

Tesch, R. (1990). *Qualitative Research: Analysis Types and Software Tools*. The Falmer Press.

Wilson, R.J. (1999). Classroom assessment investigations. *Alberta Journal of Educational Research*, 45(3),263-266.

Wilson, R.J. (2000). *A model of assessment-in-practice*. Paper presented at the annual meeting of the Canadian Society for the Study of Education, University of Alberta, May, 2000.

Wilson, R.J. & Martinussen, R.L. (1999). Factors affecting the assessment of student achievement. *Alberta Journal of Educational Research*, 45(3), 267-277.

BEST COPY AVAILABLE